WHAT DOES 5% MEAN? THE CORESET REGIME BOUNDARY IS PINNED IN SAMPLES PER CLASS

Anonymous authors
Paper under double-blind review

ABSTRACT

Coreset selection methods are usually evaluated at fixed fractions of the training set, yet the winning selection principle changes with budget. Where does this change occur, and which budget coordinate makes it stable? We measure the crossover budget m_0 between a learnability probe (LFrac) and a coverage anchor (Herding), observing the reversal directly on four datasets including ImageNet-1k. Two controlled interventions distinguish the effects of pool size and visual bandwidth under fixed selection and training protocols. Shrinking the per-class pool of ImageNet-100 by $8.4\times$ leaves m_0 within 65–80 samples per class while its fraction reading drifts by roughly $10\times$; CIFAR-100 reproduces this approximate stability. Reducing effective pixel count by $49\times$ on a fixed input canvas lowers full-data accuracy by 7.4 points, while m_0 shifts systematically from about 66 to 80 within the same sampled budget bracket. The strongest invariance is therefore to pool size: the crossover behaves like an approximately fixed per-class sample requirement of the measured selection and learning system. This is more than unit conversion, because matching fractions can place methods in different budget regimes. Locating the boundary without downstream measurement remains open: the six dataset descriptors examined here provide no validated predictor, and naive subset mixtures on CIFAR-100 do not reliably beat the better pure endpoint. Oracle-tuned CCS improves on both endpoints near the transition, but a fixed setting transferred to ImageNet-100 loses at the smallest budget. These findings motivate reporting absolute samples per class and pool size alongside fraction in cross-dataset evaluations.

1 INTRODUCTION

Coreset selection trains the downstream model on a subset of the full training set when compute, annotation, or storage is constrained, and its budget is almost always measured as a fraction f of the full set. New methods are evaluated on fixed fraction grids (Guo et al., 2022), and cross-dataset conclusions are compared at matching f (Xia et al., 2023; Zheng et al., 2023). The convention is convenient because f is dimensionless, so one grid of fractions applies to any dataset. It was adopted for that convenience rather than chosen deliberately as a coordinate, and it has not been examined as a claim in coreset evaluation.

No single selection principle wins at every budget, and this qualitative picture has both theoretical and empirical support. Theory predicts that the optimal pruning strategy flips with data volume, keeping easy samples when data is scarce and hard samples when data is abundant (Sorscher et al., 2022). Empirically, methods that keep only hard samples collapse catastrophically at high pruning rates (Zheng et al., 2023). Turning this picture into a measurable object requires a boundary that can be located. We locate it with the crossover budget m_0 between two representative selectors, both defined in §3.1. The learnability probe LFrac keeps each class's most stably learned samples under a short proxy run, and the coverage anchor Herding (Welling, 2009) is a stable reference across the full budget range. The test-accuracy difference between the two turns from positive to negative as the budget grows, and the first flip is m_0. Four datasets, including ImageNet-1k, directly exhibit the flip, and the other two stay on one side of the boundary throughout the measured range (Figure 1).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Where this boundary sits, and what moves it, have remained open questions. This paper gives controlled evidence that answering them starts with the unit in which the budget is measured. The regime boundary that governs which kind of sample to keep is stable when measured in absolute samples per class, and it drifts by nearly an order of magnitude in fraction as the per-class pool changes. Three hypotheses about the boundary are plausible a priori. It could (i) scale with the pool, keeping a constant fraction, (ii) track per-image information, because blurrier images require more images to compensate, or (iii) stay pinned at an absolute number of samples per class. Hypothesis (iii) rests on the intuition that building basic knowledge of a class takes a fixed number of samples, set by the class distribution itself. Only measurement separates the three.

Two controlled interventions on the data side separate pool-size stability from bandwidth sensitivity. Shrinking the per-class pool moves the boundary's fraction reading by nearly an order of magnitude, while its absolute reading stays in a narrow band (§4.2). Compressing every image's effective pixel count, by downsampling and then upsampling back, lowers the accuracy ceiling and produces a smaller, systematic shift within one bracket (§4.3).

Once the boundary is recognized, the natural next steps are to predict it and to exploit it, but none of the predictors or mixed selectors we tested gets around per-budget measurement. The six dataset descriptors examined here provide no validated predictor of the boundary's location, and several simple orderings fail. Naive mixtures of the two selectors do not reliably beat the better pure endpoint at any tested budget, and even the coverage-aware method CCS beats the endpoints only in the transition zone around m_0 (§4.4). Knowing which regime a budget falls in presupposes the right coordinate, and the fixed-fraction protocol systematically confounds regimes. The same $f = 0.05$ falls left of the boundary on CIFAR-100, near it on ImageNet-100, and past it on ImageNet-1k.

This paper contributes three controlled statements:

- **Phenomenon: a measurable regime boundary.** That different budgets need different selection principles no longer rests on a qualitative picture. The LFrac–Herding accuracy difference turns from positive to negative as the budget grows, the first flip locates m_0, and the reversal is directly observed on four datasets including ImageNet-1k. The win–loss pattern of the third-party method CCS is organized by the same boundary (§4.4).
- **Coordinate: the boundary is stable in absolute samples per class.** The natural coordinate for coreset budgets is absolute samples per class, $m \approx f \cdot N$, rather than fraction. Under pool shrinking the per-class reading of m_0 stays put while its fraction reading drifts by nearly an order of magnitude, and CIFAR-100 replicates the pattern (§4.2). The bandwidth intervention shifts it only modestly within one bracket (§4.3). Fraction misreads a difference of coordinates as a difference of behavior.
- **No tested substitute for measurement.** The six dataset descriptors examined here provide no validated predictor of the boundary's location. Naive mixtures of the two selectors do not reliably beat the better pure endpoint, and even granted an oracle per-budget hyperparameter sweep, CCS beats the endpoints only in the transition zone. The tuned knob carried unchanged to a second dataset keeps the transition-zone gain but loses deep below the boundary, so exploiting CCS still requires knowing the regime. Measurement-free prediction and adaptation-free exploitation both remain out of reach.

2 Related work

Existing selectors fall roughly into five families by the computation they require at selection time. Simple sampling baselines use no score at all, sampling either at random or class-balanced (Random, Uniform). Geometric prototype methods match the geometric structure of each class in feature space, with Herding as the representative (Welling, 2009; Rebuffi et al., 2017). Training-dynamics scores rank samples by statistics of the training process, such as forgetting events (Toneva et al., 2019), GraNd/EL2N (Paul et al., 2021), and Data Cartography (Swayamdipta et al., 2020). Optimization-based methods cast subset selection as an optimization of gradient matching or coverage over the full set (Mirzasoleiman et al., 2020; Killamsetty et al., 2021; Iyer et al., 2021). Coverage-aware methods no longer bet on the hardest or easiest samples alone and instead maintain coverage of the data distribution (Xia et al., 2023; Zheng et al., 2023; Maharana et al., 2024). Among these, CCS is the most directly related to the budget regimes studied here and is evaluated separately

in §4.4. All five families ask by what signal samples should be selected. This paper asks instead in what coordinate the budget itself should be measured, and in which coordinate the boundary that governs the choice of selection principle is pinned.

Conceptually closest is Sorscher et al. (2022). Their statistical-mechanics analysis of a perceptron model predicts a qualitative flip, keeping easy samples when data is scarce and hard samples when it is abundant. Their CIFAR-10 experiment varies the initial dataset size and the retained fraction jointly and shows the winning strategy flipping between the two. This paper neither repeats nor disputes that conclusion but answers a coordinate question they did not pose. The question is in what unit the boundary is stable when the candidate pool changes, and which coordinate keeps cross-dataset comparisons from confounding budget regimes. Their analysis has no class structure, measures the budget in samples per parameter, and locates the flip only by numerical solution. Read in our coordinate, their numerically solved flip line appears closer to a constant-absolute-number curve than to a constant-fraction curve. Their experiment maps win regions rather than locating a boundary under one fixed selector pair. This paper locates the boundary as a per-class number m_0 under a fixed selection and training protocol on six datasets, and tests it with two controlled interventions, pool shrinking and bandwidth, that separate three prior hypotheses. In active learning, Hacohen et al. (2022) likewise find that opposite strategies suit low and high labeling budgets, counted in absolute labeled examples. That work does not vary the candidate pool, which is the intervention that decides the coordinate question here.

3 Experimental setup

This section fixes the selectors, datasets, downstream protocol, and metrics used throughout. The experimental infrastructure follows prior work directly. The downstream training protocol and hyperparameters follow an existing coreset selection benchmark (Liu & Han, 2026), which fixed several implementation issues in DeepCore and unified the downstream test suite. Selector implementations are based on DeepCore (Guo et al., 2022).

3.1 Two selectors and the LFrac probe

The load-bearing contrast of the paper is between two selectors, the LFrac learnability probe and the Herding coverage anchor. LFrac trains a proxy network on the full training set for K epochs and, for each sample i with label y_i, records

$$c_i^{(e)} = \mathbf{1}\Big[\arg\max_c \Pr_{\theta^{(e)}}(c \mid x_i) = y_i\Big] \tag{1}$$

as the 0/1 indicator of being classified correctly after epoch e, with score

$$s_i = \frac{1}{K}\sum_{e=1}^{K} c_i^{(e)}, \tag{2}$$

the fraction of epochs in which the sample is classified correctly (hence *learnability fraction*). The score averages 0/1 indicators rather than confidences. Intuitively, samples with high s_i are the earliest and most stably learned. The per-class budget is then filled with the highest-s_i samples of each class. Scoring by short proxy training belongs to the cheap-proxy lineage (Coleman et al., 2020), and the score itself is the correctness statistic of Data Cartography (Swayamdipta et al., 2020) taken over a short proxy window. The short proxy window, per-class top-k, and localization in absolute samples per class are this paper's experimental operationalization rather than a new selection algorithm. The proxy window $K \in \{5, 10, 20, 40\}$ counts proxy training epochs, and each experiment group fixes one reporting arm (per-group listing in Appendix C, all arms in Appendix A). The crossover verdict is identical across arms with $K \geq 10$, while the numeric reading of m_0 can shift with the window, so readings are only ever compared within one arm (§4.2). Herding (Welling, 2009) greedily matches the class feature mean. Its use as a full-range reference rests on benchmark evidence showing that it never collapses catastrophically anywhere in the fraction range and stays at the level of a usable reference (Liu & Han, 2026). All other selectors appear only as secondary baselines (§3.4).

3.2 DATASETS

Six datasets span native resolutions from 32 to 224 pixels and a spectrum from high to low redundancy: ImageNet-1k, ImageNet-100, CIFAR-100, TinyImageNet, CUB-200, CIFAR-10. Per-class pools range from about 30 images (CUB-200) to 5,000 (CIFAR-10), class counts from 10 to 1,000, and granularity from the coarse object categories of CIFAR-10 to the two hundred fine-grained bird species of CUB-200.

3.3 DOWNSTREAM PROTOCOL, PAIRED SEEDS, AND FEATURE SOURCES

Training recipes and feature sources are fixed within each experiment group, and paired seeds report end-to-end reliability. The downstream model is ResNet-18 with momentum SGD, weight decay, cosine annealing, and standard augmentation. ImageNet-1k and all ImageNet-100 interventions use 90 epochs; CIFAR-10/100, TinyImageNet, and CUB-200 use 200 epochs (Appendix C). Seeds are paired, meaning the proxy seed equals the downstream seed. Each end-to-end run is then one coherent draw of (selection, initialization, training), and the reported cross-seed standard deviation measures end-to-end method reliability. The two selectors are compared within the same seed. The per-image easiness score field correlates at $r = 0.82$ across seeds, indicating a substantial shared ranking component (Appendix D). Seed counts are 3–5 per experiment group (per-group listing in Appendix C). The bandwidth experiment adds single-seed extreme budgets, and its crossover verdict relies only on the 3-seed core bracket. Feature sources for geometric selection follow one rule: every extractor is trained from scratch on the dataset it selects from. The exception is CUB-200, whose 30 images per class cannot support one, so its Herding uses frozen DINOv2 features (Oquab et al., 2024). The source is constant within each experiment group and never compared across groups (§5). Per-group sources, and two *prior-reference* anchors taken from published benchmark results rather than rerun (Herding and Uniform at $f{=}0.05$ on ImageNet-1k), are disclosed in Appendix C.

3.4 METRICS AND THE BOUNDARY DECISION RULE

This subsection defines how the boundary is measured. We report test accuracy and measure the budget in absolute samples per class, $m \approx f \cdot N$, with N the per-class pool size. This is the paper's central coordinate claim, with evidence in §4.2 and §4.3. Let $\Delta(m) = \mathrm{acc}(\mathrm{LFrac}) - \mathrm{acc}(\mathrm{Herding})$. Scanning from the low-budget side with linear interpolation between adjacent observations, we take the first positive-to-negative zero crossing as the crossover budget m_0. When the transition neighborhood shows a small back-crossing due to finite seeds, we still take the first flip and count the back-crossing toward transition-zone uncertainty. When the sign never flips within the measured range, we report a one-sided constraint on m_0. Always positive gives a right-censored lower bound, and always negative gives a left-censored upper bound, since the flip, if any, lies below the first measured budget. Other baselines (Random, class-balanced Uniform, EL2N, CCS) appear where they are needed.

4 EMPIRICAL STUDIES

This section reports four studies: the crossover boundary (§4.1), the two interventions (§4.2–4.3), and the boundary's resistance to prediction and mixing (§4.4).

4.1 THE CROSSOVER BOUNDARY AND A REDUNDANCY SPECTRUM

In absolute samples per class, LFrac and Herding exhibit a reproducible crossover budget m_0, the measurement object this paper uses to test its coordinate claim. LFrac leads on the low-budget side. The difference narrows as the budget grows and flips sign at m_0, and Herding leads beyond it. Figure 1 overviews all six datasets. This subsection establishes the boundary itself, and §4.2 and §4.3 then ask which coordinate it is pinned in.

The flip is directly observed on most datasets, and where it is observed it is not marginal. Four datasets show the sign change inside their measured range. The other two stay positive throughout and give right-censored constraints, each annotated in Figure 1. On ImageNet-1k the gap swings

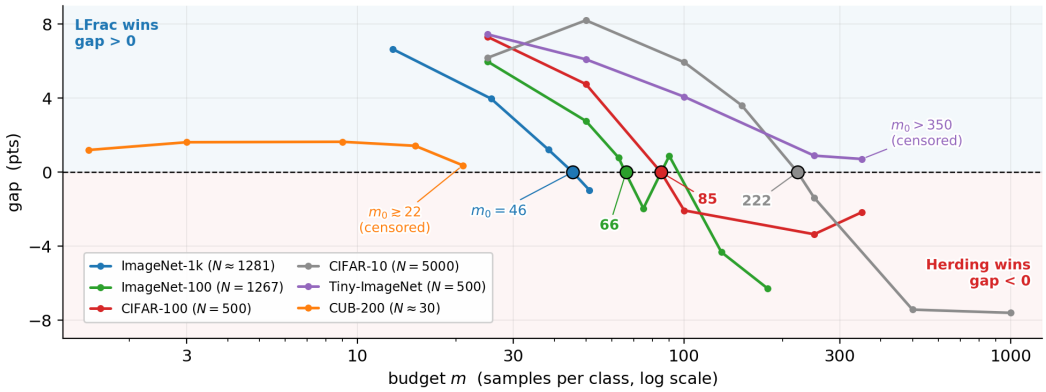


Figure 1: **The crossover in absolute samples per class, across all six datasets.** Each curve is the gap LFrac − Herding in test-accuracy points against the per-class budget m; N is the per-class pool size. Per-curve instruments and per-budget numbers are in Appendix F.

from $+6.65$ points at $m{\approx}13$ to -0.97 at $m{\approx}51$, and the CIFAR-10 swing is wider still. The boundary is therefore a large, reproducible feature of the accuracy curves rather than a marginal ordering.

The verdict does not depend on the proxy window either, although the numeric reading of m_0 can (Table 1). A longer window lets the proxy fit further before the score is read, and across arms this changes the size of the gap while leaving its sign in place in all but one case. The sign is the same in every arm with $K \geq 10$ at every budget, and $K = 5$ agrees except on CUB-200 at $f \geq 0.3$, where a 5-epoch proxy underfits the fine-grained task (Appendix A). The gap itself moves by up to 3.4 points across arms, so m_0 is only ever read within one arm. The two arms of the bandwidth experiment (§4.3) differ by less than one point everywhere and give the same crossover bracket at every bandwidth (Appendix A). The window is thus part of the instrument, not of the boundary.

	CIFAR-10			CIFAR-100			TinyImageNet			ImageNet-1k	
m per class	250	500	1000	25	50	100	25	50	100	38	51
$K = 5$	-2.8	-8.9	-9.8	$+6.0$	$+3.2$	-4.1	$+6.5$	$+3.5$	$+0.9$	$+1.1$	-0.7
$K = 10$	-2.1	-8.1	-8.2	$+6.5$	$+4.6$	-2.0	$+7.5$	$+5.3$	$+2.7$	$+1.2$	-1.0
$K = 20$	-1.5	-8.1	-7.9	$+6.2$	$+4.5$	-1.7	$+7.4$	$+6.1$	$+4.0$	$+0.7$	-1.2
$K = 30$	-1.2	-7.5	-7.6	$+7.0$	$+4.9$	-1.8	$+7.5$	$+6.1$	$+4.3$	–	–
$K = 40$	-1.4	-7.4	-7.6	$+7.3$	$+4.8$	-2.1	$+7.5$	$+6.1$	$+4.1$	–	–

Table 1: **The sign of the gap does not depend on the proxy window.** LFrac − Herding in test-accuracy points for every K arm at the budgets shared by all arms (5 seeds; ImageNet-1k 3 seeds, 3 arms). CIFAR-10's grid lies right of its crossover. The gap moves by up to 3.4 points across arms ($K = 5$; 1.6 among $K \geq 10$), and the $K = 30$ and $K = 40$ arms are within 0.3 points of each other everywhere. Means and standard deviations are in Appendix A.

The six verdicts line up qualitatively along an empirical redundancy spectrum. The organizing variable is roughly per-class redundancy, the joint effect of per-class pool size and class granularity, read as the share of the pool that the learnability regime occupies. ImageNet-1k and CIFAR-10, the most redundant pools, exhaust LFrac's advantage within 4–5% of their pools. On CIFAR-10 that small share is nonetheless $m_0 \approx 222$ samples per class, the deepest boundary directly measured. CUB-200, with 200 fine-grained classes of about 30 images each, keeps learnable samples ahead through at least 70% of its pool, and TinyImageNet does the same. This share is a readout of redundancy rather than a budget coordinate, since §4.2 shows that changing N alone moves the share while m_0 stays put. Simple directional intuition also fails at the extremes: CIFAR-10 and TinyImageNet have similar image sizes (32 and 64 pixels) yet sit at opposite ends. Whether the precise position of m_0 can be predicted from dataset statistics is tested head-on in §4.4.

N	m_0 (first)	95% CI	m_0 (last)	f_0
ImageNet-100				
1267	66.4	[62.1, 69.4]	96.8	0.052
600	71.9	[66.6, 77.6]	71.9	0.120
300	64.6	[63.0, 64.9]	64.6	0.215
150	80.4	[60.3, 88.1]	80.4	0.536
$\gamma = -0.07$ (first) / $+0.10$ (last)				
CIFAR-100				
500	43.2	[41.4, 44.7]	57.2	0.086
250	39.5	[37.9, 47.0]	61.5	0.158
150	42.6	[36.7, 47.0]	56.7	0.284
100	36.3	[28.3, 51.9]	47.6	0.363
$\gamma = +0.08$ (first) / $+0.10$ (last)				

Table 2: **Pool shrinking: crossover readings in samples per class.** First-crossing m_0 (§3.4) with seed-level bootstrap 95% interval, the stricter last-crossing diagnostic, and the fraction reading $f_0 = m_0/N$. γ is the log-log slope of m_0 on N: a fixed fraction predicts 1, a fixed per-class number 0.

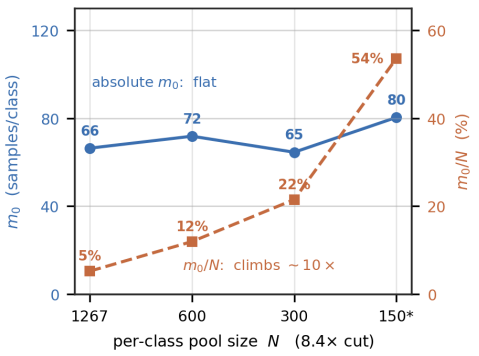


Figure 2: **Pool shrinking on ImageNet-100, dual view.** Each point is the interpolated crossover of the matching pool curve in Figure 3a (numbers in Appendix F). *N=150 is a small-pool stress test.

In summary, the two selectors are regime specialists. Which side of m_0 a budget falls on decides which principle to use, and neither selector wins across all budgets. Since the boundary is real and reproducible across datasets, the next question is which coordinate it lives in.

4.2 Pool shrinking: m_0 is pinned in samples per class, not in fraction

This subsection gives the first controlled argument for the coordinate claim: shrinking only the per-class pool leaves m_0 nearly fixed in absolute samples per class while the fraction reading drifts by nearly an order of magnitude. The intervention changes only the pool size N, and each of the introduction's three prior hypotheses then makes a definite prediction. If the boundary is pinned in fraction (i), m_0/N stays constant and m_0 must shrink in proportion to the pool. If it is pinned in absolute samples per class (iii), m_0 stays put and the fraction reading drifts passively. Hypothesis (ii), that the boundary tracks per-image information, is blind to pool size and is examined in §4.3.

Shrinking the pool leaves the crossover where it was. On ImageNet-100 we cut the per-class pool in nested steps from the full $N = 1267$ down to 150, an $8.4\times$ reduction, re-select and re-train at each level under the same protocol, and scan the absolute per-class budget m (Figure 3a). If the boundary were pinned in fraction, the smallest pool would flip at an m_0 shrunk by the same factor. Instead the four measured m_0 stay within 65–80 samples per class, while the fraction reading $f_0 = m_0/N$ climbs from 0.05 to 0.54, a drift of roughly $10\times$ (Table 2, with a seed-level bootstrap confirming the band in Appendix F). A log-log fit of m_0 against N gives an exponent near zero where a fixed fraction predicts one (γ in Table 2). A stricter last-crossing rule that reads the back-crossing of the full pool as unresolved leaves the exponent an order of magnitude below the fraction prediction. Hypothesis (i) is directly falsified, and the data are consistent with (iii).

CIFAR-100 reproduces the pattern under the same protocol. Its per-class pool is cut $5\times$ in four nested levels under the feature-source rule of §3.3. The four m_0 again stay in a narrow band, 36–43 samples per class, while the fraction reading climbs from 0.09 to 0.36, a drift of roughly $4\times$ (Table 2). Hypothesis (i) is falsified a second time, and the data are again consistent with (iii). Herding seed variance is large near the transition, so we report a band, not a point estimate.

This band sits below the full-pool CIFAR-100 reading of §4.1, and the difference is an instrument shift, not a contradiction. The two groups measure the same pool but differ in Herding feature source and budget grid, not merely in proxy window (§3.3): within §4.1's own group, swapping the window from K=40 to K=10 leaves the full-pool reading near 85 either way (Appendix A). The two instruments agree that the flip exists and disagree only on its numeric reading, so each pool-shrinking verdict is read within one instrument.

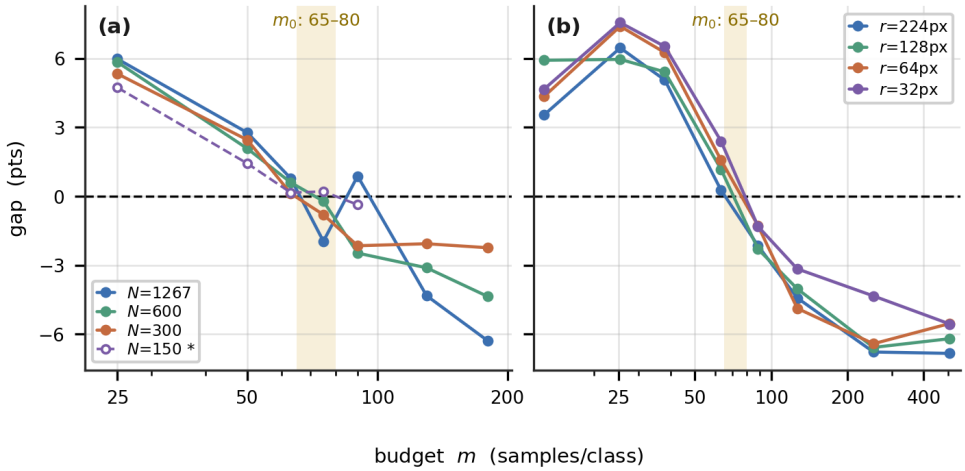


Figure 3: **The crossover band holds under both interventions.** (a) Pool shrinking (§4.2); (b) fixed-canvas bandwidth (§4.3). The full-pool/native reference (blue) is one system measured on each ladder's own budget grid, so its sampled points differ between panels; its crossover does not. *N=150: small-pool stress test.

Three checks rule out a pool artifact. The first targets convergence: a small pool could force the two selectors onto the same samples, and such an artifact would pin the flip to a fixed overlap level, a fixed m/N. It does not. On the ImageNet-100 pools the selectors' overlap at the crossover runs from 0.06 on the full pool to about 0.5 at N=150 while m_0 stays in the band (Appendix F). The second targets pool quality: if a smaller pool held worse samples, a fixed absolute budget drawn from it would train worse and need a larger fraction to compensate. A class-balanced random check on the CIFAR-100 pools holds the training set at $m = 40$ random samples per class and varies only the source pool, and the accuracies stay within 0.9 points of one another across the four pools (Appendix F). The third targets sample identity, the complementary story that the band survives because the selected samples themselves survive. They do not. On the ImageNet-100 pools LFrac selections change identity almost completely across levels while the easiness score fields remain strongly correlated, and the band holds regardless (Appendix D). What a selection does is set by its composition ratio, the share of easy versus hard samples it keeps, not by which samples carry it.

In summary, fraction misreads a difference of coordinates as a difference of behavior. On the same dataset, merely shrinking the pool moves the same fraction from one side of the boundary to the other. The fraction $f = 0.10$ is already past m_0 on full-pool ImageNet-100 ($m = 127$) but still far to its left on the $N = 150$ pool ($m = 15$). Figure 2 gives the dual view, and complete curves for the four pools are in Appendix F.

4.3 Visual bandwidth: the ceiling and the boundary dissociate

The bandwidth intervention measures how the boundary responds when each image is degraded while the input canvas and learner stay fixed. If blur removes useful information, more samples might be needed and m_0 might move right. Unlike the constant-fraction hypothesis, this account gives no quantitative prediction without a model relating image information to sample requirements.

The intervention changes only the effective resolution. Every ImageNet-100 image is downsampled to $r \times r$ ($r \in \{32, 64, 128, 224\}$) and upsampled back to 224×224. The tensor size, the ResNet-18, and the training protocol are unchanged, so the only variable is the effective pixel count per image, which differs by $(224/32)^2 = 49\times$ between $r = 32$ and $r = 224$.

The accuracy ceiling and the crossover respond on different scales. Removing pixels lowers the full-data ceiling but moves the boundary only within the bracket that all four resolutions share. Full-data accuracy falls monotonically by 7.4 points from r=224 to r=32, while all four mean gap curves cross between the same two adjacent budgets of the grid (Figure 3b; numbers in Appendix F). That

bracket is the sampled budget grid, not a confidence interval. Within it, the interpolated m_0 shifts monotonically from about 66 to 80 as resolution falls, a rise of about $1.2\times$ whose direction a paired seed-level bootstrap supports. Bandwidth thus leaves the boundary inside one bracket while moving the ceiling well beyond seed noise: a modest but systematic response, not zero.

The two interventions support different conclusions. Pool shrinking rejects a constant-fraction boundary and supports approximate stability in absolute samples per class. Bandwidth reduction produces a smaller within-bracket shift alongside a substantial ceiling change (Figure 3b). Because pixel count is not a calibrated measure of classification-relevant information, the bandwidth result does not falsify hypothesis (ii) or distinguish it from a change in class-distribution complexity.

4.4 Neither prediction nor mixing substitutes for measuring the boundary

Two natural next steps follow from recognizing the boundary: predicting its location, which would spare per-dataset downstream measurement, and exploiting it in a stronger mixed method that captures both sides' advantages. The descriptor checks yield no validated prediction rule. Naive mixing fails, while oracle-tuned stratified sampling wins only in the transition zone.

The six dataset descriptors examined here do not establish a reliable rule for locating m_0. Because censored m_0 are not ordinary point estimates, we ask whether a descriptor can explain or rank the observed and censored constraints rather than regressing on m_0. The candidates are total pixels per class, per-class effective rank of DINOv2 features, proxy training dynamics, proxy fitting difficulty, image resolution, and inter-class confusion (definitions in Appendix E). Four fail a contrast drawn within a single instrument, and the other two give no stable ordering at all. Total pixels per class and proxy fitting difficulty point the wrong way, since ImageNet-100 is easier for the proxy to fit than ImageNet-1k yet has the larger m_0, about 66 against 46. Image resolution cannot separate the same pair, both at 224 pixels. Per-class effective rank reverses within one instrument: CIFAR-10 has the highest rank yet a boundary of 222, below TinyImageNet's censored bound, and the reversal holds at every sample count tried. Inter-class confusion gives no stable direction, because which of the two ImageNets counts as more confused depends on how the confusion entropy is normalized for class count. Proxy training statistics vary with fitting and memorization, and CUB-200's censored boundary cannot anchor an ordering against the observed crossovers. No descriptor tested here can replace a downstream measurement of the boundary.

No naive subset mixture reliably beats the better pure endpoint at any tested budget. To test whether the two principles contribute complementary samples, on CIFAR-100 we splice an LFrac easy core, a share α of the budget, with a tail chosen by another principle, and the three constructions differ only in the tail. Construction A runs Herding on the leftover pool with its greedy mean initialized at the core, so the tail repairs the core's residual to the class centers. Construction B runs Herding on the leftover pool unaware of the core, and construction EL2N takes the leftover pool's hardest samples by EL2N score. Across six budgets spanning the crossover, the best interior blend minus the better pure endpoint ranges from -4.67 to $+0.23$ points, and the only two positive values, both from construction B, are within 3-seed noise. Most interior blends are significantly worse. Construction A's full α profile is a valley rather than an interpolation, and at three of the six budgets every interior blend falls below both endpoints (Figure 4; budget grid and full per-budget tables in Appendix D). A fixed blend does not average the two regimes, so a naive blend cannot replace knowing which side of the boundary a budget lies on.

Even the coverage-aware CCS (Zheng et al., 2023), evaluated separately under a protocol that favors it, has its wins and losses organized by the boundary. CCS samples randomly within difficulty strata and excludes the hardest β share, where β is a hyperparameter tuned per pruning rate. We port the official implementation, take the difficulty score as the same proxy's correctness mean, sweep β over the full feasible grid, and score CCS at each budget by its best β (grid details in Appendix D). This is an oracle protocol, since the pure endpoints have no tunable knob and a practitioner cannot know β without a round of downstream measurement. Under this most favorable reading the boundary splits CCS's record into three segments (Table 3).

Net gains appear only in the transition zone around m_0. There the two pure endpoints differ by less than 0.6 points, and deep on either side the best CCS merely ties the correct endpoint for that side. The best β falls as the budget grows. A larger β discards more of the hardest samples and keeps an easier subset, so this trend is the same easy-to-coverage flip read through CCS's knob. The knob

Segment	f	m	Best β	Gain
Left of boundary	0.10	50	0.6	+0.75
	0.12	60	0.5	+1.03
Transition zone	0.14	70	0.3	+2.97
	0.17	85	0.4	+3.18
Right of boundary	0.20	100	0.2	+0.24
	0.25	125	0.3	+0.07

Table 3: **CCS on the full feasible β grid (CIFAR-100, 3 seeds).** Gain is the best per-budget β minus the better pure endpoint, in points. Only the transition-zone gains exceed 3-seed noise.

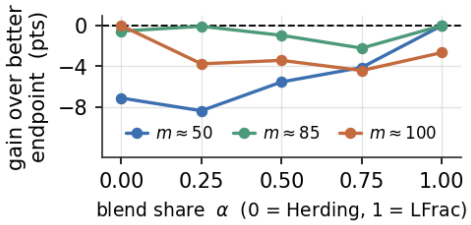


Figure 4: **Naive mixing does not beat the better pure endpoint (CIFAR-100).** Gain over the better endpoint against the blend share α, construction A; constructions B and EL2N agree (Appendix D).

does the main work: at f=0.10, moving β from 0 to its optimum is worth 16.2 points. This result delimits the section's negative conclusion rather than overturning it.

One escape would remain if the sweep could be run once and carried across datasets, and the CIFAR-100 grid even nominates a candidate. A fixed β=0.4 never loses to the better endpoint beyond seed noise at any of its six budgets. Carried to ImageNet-100 with no tuning, it reproduces the three-segment shape but not the never-loses property. At m=25 it loses to the better endpoint by 3.4 points, then ties it at the next budget, gains through the transition zone, and returns to a tie at the largest budget (Appendix D). The failure sits exactly where regime knowledge was missing: a fixed β knows nothing about the next dataset's boundary, and m=25 lies deep below m_0 in a range the CIFAR-100 grid never reached. Skipping the sweep is safe only when the regime is already known.

The tested descriptors provide no validated prediction rule, and CCS's carried setting loses its never-loses property on transfer. These experiments leave open whether a validated shortcut to boundary measurement exists, which motivates a stable coordinate for that measurement.

5 Discussion and Conclusion

The crossover between a learnability probe and a coverage anchor is a measurable boundary, observed directly on four datasets. Under pool shrinking its per-class reading stays put while the fraction reading drifts by nearly an order of magnitude. The bandwidth intervention moves the ceiling substantially and the boundary only modestly. The natural coordinate for coreset budgets is therefore absolute samples per class, $m \approx f \cdot N$. No tested descriptor, mixture, or transferred CCS setting removes the need to measure the boundary per dataset, so its coordinate must be stable.

For practice, fixed-fraction cross-dataset evaluation systematically compares different budget regimes. Judged by each dataset's own measured curves, the same $f = 0.05$ sits left of the crossover on CIFAR-100, near it on ImageNet-100, and past it on ImageNet-1k. Within-dataset comparisons at the same f are unaffected. Cross-dataset comparisons, budget sweeps, and benchmark grids should align points in absolute samples per class m, and reports should give m and N alongside f. Dataset distillation counts images per class (IPC) and few-shot learning counts shots (k-shot), so coreset selection would return to the native unit of its neighbors.

The coordinate also has a statistical motivation. Approximating a class-conditional distribution to a given fidelity depends on its complexity, not on the size of a pool drawn from it. Theoretical coreset bounds likewise depend on complexity parameters, not on full-set size (Feldman & Langberg, 2011), consistent with the pool-shrinking result and the near-constant Uniform accuracy at fixed m.

Limitations. The numeric value of m_0 depends on the selector pair and the feature source, so the paper establishes the invariance of the boundary within one measuring instrument, not an instrument-independent constant. All interventions act on the data side and the downstream model is a single ResNet-18, so the evidence locates the boundary and its coordinate but does not identify its cause. The mixing study is confined to CIFAR-100.

ETHICS STATEMENT

All experiments use public image classification datasets, including ImageNet-1k and CUB-200, under their original research licenses. The work involves no human subjects, personal data, or new data collection.

REPRODUCIBILITY STATEMENT

The main-text numbers of every experiment group were re-aggregated from raw result files and checked against reference values, with zero verification failures. The ImageNet-1k Herding and Uniform anchors (prior-reference) and the descriptor statistics of §4.4 are outside the table's coverage and are marked separately where they appear. Appendix C specifies the selection and training protocol and rounding rules, and Appendix F lists the numbers underlying every main-text figure. The per-run result files, selected-index files, and analysis scripts will be released with the paper.

Experiment group	Datasets	Result files
Cross-dataset phenomenon (5 seeds)	CIFAR-10/100, TinyImageNet	315
Low-budget crossover	ImageNet-1k	78
Low-budget extension (5 seeds)	CIFAR-10	40
Fine-grained phenomenon	CUB-200	60
Nested pool shrinking	ImageNet-100	234
Pool-shrinking replication	CIFAR-100	297
Fixed-canvas visual bandwidth	ImageNet-100	268
Mixing anti-synergy + CCS (full β grid)	CIFAR-100	387
CCS fixed-β transfer	ImageNet-100	18
Wide budget sweep (f=0.01–0.70, 5 seeds)	CIFAR-100, TinyImageNet	150
Total		1847

AI USE STATEMENT

In this work, we used generative AI tools to implement standard components of the experimental pipeline, to assist with data analysis, and to draft and edit parts of the paper. We have reviewed all AI-assisted work: the generated code was checked against the reference implementations it was adapted from and rerun by the authors, and all drafted and edited text was read and approved by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. In *International Conference on Learning Representations (ICLR)*, 2020.

Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *ACM Symposium on Theory of Computing (STOC)*, 2011.

Chengcheng Guo, Bo Zhao, and Yanbing Bai. Deepcore: A comprehensive library for coreset selection in deep learning. In *International Conference on Database and Expert Systems Applications (DEXA)*, 2022.

Guy Hacohen, Avihu Dekel, and Daphna Weinshall. Active learning on a budget: Opposite strategies suit high and low budgets. In *International Conference on Machine Learning (ICML)*, 2022.

Rishabh Iyer, Ninad Khargonkar, Jeff Bilmes, and Himanshu Asnani. Submodular combinatorial information measures with applications in machine learning. In *Algorithmic Learning Theory (ALT)*, 2021.

Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. GLISTER: Generalization based data subset selection for efficient and robust learning. In *AAAI Conference on Artificial Intelligence*, 2021.

Yangze Liu and Zhongyi Han. Are coreset selection methods worth their cost? *arXiv preprint arXiv:2609.22894*, 2026. URL https://arxiv.org/abs/2609.22894.

Adyasha Maharana, Prateek Yadav, and Mohit Bansal. D2 pruning: Message passing for balancing diversity and difficulty in data pruning. In *International Conference on Learning Representations (ICLR)*, 2024.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning (ICML)*, 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: Beating power law scaling via data pruning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations (ICLR)*, 2019.

Max Welling. Herding dynamical weights to learn. In *International Conference on Machine Learning (ICML)*, 2009.

Xiaobo Xia, Jiale Liu, Jun Yu, Xu Shen, Bo Han, and Tongliang Liu. Moderate coreset: A universal method of data selection for real-world data-efficient deep learning. In *International Conference on Learning Representations (ICLR)*, 2023.

Haizhong Zheng, Rui Liu, Fan Lai, and Atul Prakash. Coverage-centric coreset selection for high pruning rates. In *International Conference on Learning Representations (ICLR)*, 2023.

A SENSITIVITY TO THE PROXY WINDOW K

This section reports the complete numbers for all K arms and checks the sensitivity of the crossover verdict to K. Conclusion: the crossover verdict (the sign relative to Herding and the flip bracket) is consistent across all arms with $K \geq 10$; $K = 5$ is likewise consistent in the low-budget regime on ImageNet-1k, and turns negative only on the fine-grained CUB-200 at $f \geq 0.3$. The sensitivity of absolute accuracy to K concentrates in "shortest window $\times$ higher budget": the $K = 5$ arm on Tiny-ImageNet is 2.6/3.2 percentage points below $K = 40$ at f=0.10/0.20, but the sign is unchanged. All standard deviations are cross-seed sample standard deviations (ddof=1), throughout.

Phase D five-arm K-sweep (5 seeds). The verdict is consistent arm by arm across the three datasets: TinyImageNet all positive, CIFAR-10 all negative on this grid (whose budgets all sit right of its crossover; see the low-budget extension below), CIFAR-100 flipping between f=0.10 (positive) and f=0.20 (negative) in every arm.

CIFAR-10:

Arm	f0.05	f0.10	f0.20
$K = 5$	61.83±2.63	68.87±2.01	77.13±1.67
$K = 10$	62.48±1.46	69.65±1.69	78.76±1.27
$K = 20$	63.12±1.46	69.67±2.04	79.05±1.00
$K = 30$	63.43±1.49	70.29±1.83	79.39±1.04
$K = 40$ (main-table arm)	63.22±1.32	70.31±1.88	79.36±0.83
Herding	64.60±1.70	77.74±1.81	86.97±0.26

CIFAR-10 low-budget extension ($K = 40$ arm, 5 seeds). The phase-D grid starts at $m = 250$ per class (f=0.05) and is negative throughout. Extending the same instrument down to $m = 25$–150 shows the positive side the grid missed: the spliced series flips between m=150 and m=250, with interpolated $m_0 \approx 222$. The earlier "negative throughout" reading was left-censoring by the grid, not a boundary at zero, and CIFAR-10 enters Figure 1 as a directly observed crossover.

m per class	25	50	100	150
LFrac (K=40)	38.00±2.06	45.85±0.84	53.16±1.38	56.46±1.69
Herding	31.82±2.09	37.64±1.51	47.22±2.51	52.85±2.68
Δ (LFrac − Herding)	+6.18	+8.21	+5.94	+3.61

CIFAR-100:

Arm	f0.05	f0.10	f0.20
$K = 5$	31.59±1.22	40.81±0.99	53.43±0.65
$K = 10$	32.03±0.51	42.16±0.93	55.58±0.72
$K = 20$	31.76±1.13	42.04±0.86	55.81±0.47
$K = 30$	32.57±0.87	42.42±0.99	55.70±0.37
$K = 40$ (main-table arm)	32.87±0.45	42.32±1.06	55.46±0.32
Herding	25.55±1.64	37.57±1.23	57.53±1.25

TinyImageNet:

Arm	f0.05	f0.10	f0.20
$K = 5$	29.79±0.58	39.74±0.38	48.20±0.29
$K = 10$	30.80±0.41	41.52±0.27	50.07±0.14
$K = 20$	30.78±0.42	42.31±0.29	51.36±0.26
$K = 30$	30.82±0.53	42.36±0.42	51.62±0.34
$K = 40$ (main-table arm)	30.79±0.59	42.33±0.20	51.40±0.20
Herding	23.34±0.29	36.23±0.86	47.33±0.35

ImageNet-1k three arms (3 seeds). Every arm is positive at f=0.03 and negative at f=0.04; the flip bracket is arm-independent, and the main text reports the fixed $K = 10$ arm. The Herding and Uniform values at f=0.05 are prior-reference anchors from an existing benchmark (44.98 / 41.88), not rerun in this project.

Arm	f0.01	f0.02	f0.03	f0.04	f0.05
$K = 5$	18.04±0.48	30.40±0.23	36.69±0.23	40.44±0.07	–
$K = 10$	18.70±0.53	30.58±0.37	36.82±0.08	40.18±0.14	42.71±0.13
$K = 20$	18.45±0.22	30.33±0.24	36.29±0.16	39.93±0.09	42.28±0.06
Herding	12.05±0.39	26.61±0.31	35.61±0.22	41.15±0.20	44.98†
Uniform	9.17±0.29	21.12±0.23	29.95±0.20	36.35±0.38	41.88†

† prior-reference (§3.3).

CUB-200 three arms (3 seeds). $K = 10/20$ are positive throughout; $K = 5$ turns negative at $f \geq 0.3$. The cause is that the probe itself degenerates: on the fine-grained task a 5-epoch proxy underfits severely, and ranking noise drowns the signal. This is the only exception in all experiments where the crossover verdict is sensitive to K, and it involves only the shortest window. The Herding column is our own measurement: deterministic greedy herding on frozen DINOv2 features (§3.3), the same per-class budget rule as the LFrac arms, 3 downstream seeds.

Budget (approx. m per class)	$K = 5$	$K = 10$	$K = 20$	Herding
f0.05 (2)	4.08±0.17	4.62±0.16	4.41±0.24	3.42±0.12
f0.10 (3)	5.81±0.07	6.41±0.17	6.09±0.41	4.79±0.13
f0.30 (9)	13.75±0.48	16.20±0.35	17.09±0.47	14.56±0.40
f0.50 (15)	22.92±1.25	24.54±0.50	25.94±0.58	23.11±0.08
f0.70 (21)	34.18±0.07	35.55±0.33	35.49±0.63	35.19±0.83

Bandwidth experiment, two arms. $K = 10$ (the reporting arm) and $K = 20$ differ by less than 1 percentage point at every one of the 32 points (4 bandwidths × 8 budgets), with the same crossover bracket at every bandwidth (numbers in Appendix F.3).

B WIDE-BUDGET SWEEP ($f = 0.01$–0.70)

This section delivers the wide-budget sweep cited in §4.1 and explains the arm selection. The sweep uses a single arm across 9 budgets: the native $K = 20$ proxy (cosine learning rate fully decayed within 20 epochs), 5 seeds. The main-table range in the main text (f=0.05–0.20) cites the $K = 40$ main-table arm from Phase D (Appendix A); the two arms differ by no more than 0.75 percentage points in paired comparison on shared budgets, with identical sign and verdict everywhere. Historically, a truncation of "the first 20 epochs of a long proxy" was used as the $K = 20$ score; its paired difference against the native 20-epoch proxy is ≤ 0.45 percentage points point-by-point (within seed noise), and this paper always reports the native arm.

CIFAR-100: the flip bracket agrees with the §4.1 verdict (f=0.10 positive, f=0.20 negative), and after the flip the sign does not return to positive through f=0.70.

f	m (per class)	LFrac ($K = 20$)	Herding	Diff
0.01	5	15.29±0.68	10.76±0.71	+4.53
0.02	10	21.16±0.59	15.82±0.31	+5.35
0.03	15	25.65±0.69	18.82±0.83	+6.83
0.05	25	32.45±0.39	25.55±1.64	+6.89
0.10	50	41.58±1.15	37.57±1.23	+4.01
0.20	100	55.73±0.48	57.53±1.25	-1.80
0.30	150	61.63±0.36	64.83±0.86	-3.19
0.50	250	68.33±0.40	71.61±0.95	-3.29
0.70	350	73.48±0.22	75.72±0.24	-2.24

TinyImageNet: the sign remains positive all the way to $f = 0.70$ (350 samples per class), giving the right-censored constraint $m_0 > 350$ in the main text.

f	m (per class)	LFrac ($K = 20$)	Herding	Diff
0.01	5	10.70±0.64	6.18±0.29	+4.51
0.02	10	19.05±0.85	11.86±0.90	+7.19
0.03	15	24.68±0.79	16.81±0.98	+7.87
0.05	25	30.66±0.31	23.34±0.29	+7.33
0.10	50	41.78±0.17	36.23±0.86	+5.55
0.20	100	50.78±0.22	47.33±0.35	+3.45
0.30	150	54.89±0.34	52.98±0.32	+1.91
0.50	250	60.06±0.17	59.43±0.36	+0.64
0.70	350	63.58±0.26	63.21±0.29	+0.37

Two honest clarifications. First, at the lowest-budget end (5–15 samples per class) the accuracy gap first widens with budget and then narrows: in the extreme-scarcity regime both selectors are near failure, and the main text's "the accuracy gap narrows as budget grows" describes the main range in the neighborhood of m_0. Second, the extreme budgets at both ends of the bandwidth experiment (§4.3) (f=0.01/0.02/0.20/0.40) are single-seed exploratory points that do not participate in the crossover verdict; the verdict relies only on the 3-seed bracket over f=0.03–0.10, with complete numbers in Appendix F.3.

C Implementation details and rounding rules

Per-class budget rounding. LFrac selects offline, per class, exactly the m samples with the highest s_i (PreSelected indices), exactly m per class. Herding and Uniform use the native DeepCore implementation with class balancing on, taking fraction m/N, likewise exactly m per class.

Native proxy window. The K-epoch arm is produced by an independently trained K-epoch proxy, with the cosine learning rate fully decayed within K; we do not truncate the first K epochs of a longer proxy. The two approaches differ by ≤ 0.45 percentage points in paired comparison (Appendix B); we adopt the native arm for protocol cleanliness, not for performance.

Reporting arm per experiment group. The table below lists the reporting arm used by each experiment group.

Experiment group	Reporting arm
Cross-dataset phenomenon (§4.1)	$K = 40$
Wide-budget sweep (Appendix B)	$K = 20$ (native)
ImageNet-1k crossover	$K = 10$ (all arms $K \in \{5, 10, 20\}$ in Appendix A)
CUB-200	$K = 10$ (all arms $K \in \{5, 10, 20\}$ in Appendix A)
ImageNet-100 pool shrinking / bandwidth (§4.2–4.3)	$K = 10$; bandwidth adds a $K = 20$ arm
CIFAR-100 pool-shrinking replication (§4.2)	$K = 10$
Mixing experiment easy core (§4.4)	$K = 20$

Removal of $m = 105$. The ImageNet-100 pool-shrinking grid originally included $m = 105$: 105 samples per class total 10500 samples, which divided by batch size 256 leaves a remainder of 4, producing a final batch of 4 samples each epoch. Under `drop_last=False` this final batch contaminates BatchNorm's running statistics, randomly causing some runs to collapse by about 10 percentage points early in training (evidence: anomalous training accuracy at epoch 9, and the best test accuracy breaking monotonicity). The final batch at every other grid point is ≥ 40 samples, in the safe zone. We remove $m = 105$ rather than change `drop_last`, because the latter would cause the smallest-budget point ($m = 25$, 2500 samples total) to lose about 10% of its samples.

Tie handling at the per-class cut. With K=10 the LFrac score takes at most 11 distinct values, so ties at the top-m cut are common. Ties are broken by the deterministic order of the sorting routine, reproducibly but without a principled rule; no randomization is applied, and the churn of list membership near the cut across seeds is reported in F.2b.

Downstream training duration. CIFAR-10/100 (including the CIFAR-100 pool controls), TinyImageNet, and CUB-200 use 200 epochs. ImageNet-1k and all ImageNet-100 groups (pool shrinking, fixed-canvas bandwidth, and fixed-β transfer) use 90 epochs. The epoch count is fixed across budgets within each group.

Seeds and feature source. The paired protocol in which the proxy seed equals the downstream seed runs through all experiment groups (§3.3). Seed counts: 5 for the cross-dataset phenomenon experiments (CIFAR-10, CIFAR-100, TinyImageNet); 3 for ImageNet-1k, CUB-200, pool shrinking, and mixing; the bandwidth experiment uses 3 seeds on the core bracket (f=0.03–0.10) and 1 on the extreme budgets at both ends.

Feature sources per group. On CIFAR-10, CIFAR-100, and TinyImageNet, Herding uses the penultimate-layer features of the same 40-epoch proxy that supplies the LFrac score, under the same paired seed; the ImageNet-1k crossover and the controlled interventions of §4.2–§4.3 use a native ResNet-18 extractor trained briefly on the corresponding data. CUB-200 is the one dataset whose 30-image classes cannot support training an extractor in place: its Herding uses frozen DINOv2 features (dinov2_vits14; Oquab et al., 2024), the strongest off-the-shelf choice for the coverage anchor. The sources never serve as evidence for the same controlled variable. Two anchors are taken from published benchmark results rather than rerun, marked *prior-reference*: Herding and Uniform at f=0.05 on ImageNet-1k. They extend that series one point past the flip; the flip bracket itself (f=0.03–0.04) and all other numbers are measured in this project.

D FULL TABLES FOR THE MIXING EXPERIMENT

Definition of GAIN. For each budget, GAIN = accuracy of the best interior mixture − accuracy of the better of the two pure endpoints. GAIN ≤ 0 means the mixture produces no synergy. Across the six budgets $f \in \{0.10, 0.12, 0.14, 0.17, 0.20, 0.25\}$ (about 50–125 samples per class, 3 seeds), GAIN is $-4.13/-1.82/-0.96/-0.07/-3.43/-2.01$ for construction A, $-1.73/-1.24/+0.18/+0.23/-2.14/-1.82$ for construction B (the two positive values within 3-seed noise), and $-4.67/-3.52/-3.21/-2.77/-2.23/-1.42$ for EL2N. Figure 4 plots construction A's full α profile. The pure-endpoint columns of these tables also reproduce the crossover independently: LFrac wins at f=0.10, near-ties at f=0.17, and Herding wins from f=0.20. The tables below give the full α grid (3 seeds, where α is the budget share allocated to the LFrac easy core, $\alpha = 0$ is the pure other endpoint, and $\alpha = 1$ is pure LFrac).

Construction A (Herding residual tail):

α	f0.10	f0.12	f0.14	f0.17	f0.20	f0.25
0.00	35.64±0.23	45.56±0.84	50.62±1.55	53.22±1.66	58.51±0.78	62.54±0.44
0.25	34.39±0.78	46.36±0.61	49.66±1.29	53.70±2.03	54.74±0.33	60.53±2.12
0.50	37.23±0.56	45.02±1.57	47.41±1.73	52.80±0.91	55.08±1.05	59.15±0.47
0.75	38.65±0.53	46.05±1.55	48.19±0.91	51.54±2.10	54.08±1.16	59.34±1.44
1.00	42.78±1.79	48.18±1.17	50.17±1.30	53.76±0.14	55.84±0.29	59.76±0.41

Construction B (original Herding margin tail):

α	f0.10	f0.12	f0.14	f0.17	f0.20	f0.25
0.00	35.64±0.23	45.56±0.84	50.62±1.55	53.22±1.66	58.51±0.78	62.54±0.44
0.25	37.86±0.18	44.08±0.95	50.12±1.31	53.99±1.27	55.69±2.11	60.72±1.15
0.50	39.15±1.95	45.97±1.51	50.51±2.07	53.69±1.41	56.37±0.20	59.38±0.26
0.75	41.06±0.58	46.94±1.37	50.80±0.66	53.46±2.11	55.62±0.79	60.10±0.64
1.00	42.78±1.79	48.18±1.17	50.17±1.30	53.76±0.14	55.84±0.29	59.76±0.41

Construction EL2N (EL2N hard tail):

α	f0.10	f0.12	f0.14	f0.17	f0.20	f0.25
0.00	21.81±1.38	33.13±0.23	39.07±1.45	43.27±2.11	47.24±1.84	56.17±2.27
0.25	29.87±0.51	36.51±1.59	41.49±1.40	48.39±0.66	49.79±4.10	58.18±1.13
0.50	33.87±1.38	41.71±1.13	44.56±0.37	49.69±1.36	51.89±0.98	57.25±0.74
0.75	38.12±0.49	44.66±1.09	46.96±0.34	50.99±1.81	53.61±1.02	58.34±0.83
1.00	42.78±1.79	48.18±1.17	50.17±1.30	53.76±0.14	55.84±0.29	59.76±0.41

CCS (Zheng et al., 2023; port of the official algorithm). The difficulty score is the mean correctness of the same proxy; β is the share of the hardest samples removed, swept over the entire feasible grid in steps of 0.1 ($\beta \leq 1 - f$, infeasible entries marked –), and the main text reports the best β per budget. Same-protocol Uniform is 35.05 / 41.87 / 48.16 / 53.85 / 55.26 / 61.78 (f0.10–f0.25, same order as the table).

β	f0.10	f0.12	f0.14	f0.17	f0.20	f0.25
0.0	27.36±1.64	36.19±0.82	39.36±0.38	46.62±1.85	47.27±1.54	56.00±1.51
0.1	33.93±1.53	43.99±1.70	47.52±1.31	54.45±0.76	56.35±1.19	60.99±0.08
0.2	38.67±0.60	46.40±1.08	51.25±0.35	56.03±0.32	58.75±0.22	62.26±1.93
0.3	40.17±1.72	47.58±1.34	53.59±0.36	55.93±1.49	58.33±0.93	62.61±0.67
0.4	42.56±0.49	48.81±2.99	53.52±1.16	56.94±0.82	58.60±0.20	61.73±0.88
0.5	42.70±1.70	49.21±2.14	52.90±0.21	54.40±1.16	57.98±0.73	61.58±0.42
0.6	43.53±1.56	48.16±1.60	50.36±0.46	53.99±1.38	56.55±0.53	59.31±0.20
0.7	41.93±1.04	47.14±0.17	49.85±0.73	52.23±0.94	53.90±0.80	57.10±0.29
0.8	40.69±1.02	43.50±0.63	47.02±0.31	49.92±0.79	51.13±0.30	–
0.9	36.56±0.28	–	–	–	–	–

β carries the main work: on f=0.10, β from 0 to the best 0.6 yields +16.2. The best β decreases overall with budget (0.6 / 0.5 / 0.3 / 0.4 / 0.2 / 0.3, f0.10→f=0.25), all inside the feasible grid: at the upper end of the feasible region $\beta = 1 - f$ degenerates into taking the globally easiest $f \cdot N$ samples with no class balancing (the official implementation prunes and stratifies over the whole training set), and is clearly worse ($\beta = 0.9$ at f=0.10 is 36.56); this degenerate set is class-unbalanced and therefore not the per-class LFrac endpoint (42.78). This agrees with the CCS paper's setting that "the pruning rate must be tuned point-by-point with the coreset rate": stratified sampling itself ($\beta = 0$) lags substantially at all budgets. The table contains another regularity worth declaring: a fixed $\beta = 0.4$ does not lose to the better endpoint beyond seed noise at all six budgets (still +2.9/+3.2 in the transition zone). Recognizing this value relies on the full-grid sweep, so we tested its transfer. The same $\beta = 0.4$, carried to ImageNet-100 with no tuning (K=10 proxy correctness score, paired seeds, endpoints the same-budget LFrac and Herding runs, 3 seeds), reproduces the three-segment shape but not the "never loses" property: at m=25, the first budget deep below m_0 tested on either dataset (the CIFAR-100 grid starts at m=50, below its observed crossover), fixed-β CCS loses to LFrac by 3.43 points.

m	CCS β=0.4	LFrac	Herding	Gain over better endpoint
25	25.57±0.15	29.00±0.09	23.02±0.24	−3.43
50	38.82±0.34	38.91±0.12	36.14±1.80	−0.09
63	43.82±0.64	41.96±0.48	41.18±0.45	+1.86
75	47.13±0.49	44.16±0.49	46.12±0.47	+1.01
90	50.07±0.17	46.77±1.07	45.89±0.24	+3.31
130	56.99±0.30	52.43±0.51	56.76±0.34	+0.23

At m=90 the better endpoint is LFrac (46.77) because of the published back-crossing in the transition neighborhood (§3.4); against Herding the gain would be +4.18.

Identity overlap and functional composition. A methodological fact to state up front: low identity overlap of the subsets cannot be used to argue for or against the presence of functional redundancy. The Jaccard of top-m membership across seeds for the same selector is only about 0.15,

yet performance is unchanged within seed noise, because the underlying per-image easiness score field is highly correlated across seeds (per-class Pearson $r = 0.82$): the many tied samples near the threshold permute between seeds, and function is determined by the composition ratio, not by specific identity. Therefore the evidence for the negative mixing result is accuracy itself, not any overlap statistic. Related composition statistics: after pool shrinking, the inclusion rate of the LFrac selection relative to the full-pool selection drops from 0.34 to 0.06 as N shrinks (all above the random baseline $m/1267$), and the cross-pool easiness score field has Pearson 0.78 / 0.70 / 0.59 ($N = 600/300/150$ against the full pool); at the two ends of the bandwidth ($r = 32$ vs. $r = 224$) the normalized overlap of the selections is only 0.20–0.37, i.e. 60–80% of the selected samples differ, yet the crossover remains in the same sampled bracket, ruling out the "same batch of samples" artifact explanation (§4.3).

E DESCRIPTORS FOR THE SIX DATASETS

Each descriptor is given with its definition, computation, and the evidence available to assess it. The test criterion is whether it can explain or order the observed and censored m_0 constraints (CUB-200 $\gtrsim 22$ right-censored; ImageNet-1k ≈ 46; ImageNet-100 ≈ 66; CIFAR-100 $\approx$ 85–100; CIFAR-10 ≈ 222; TinyImageNet > 350 right-censored), not to regress on m_0. The constraints mix the instrument groups declared in §3.3 (a 10-epoch proxy window for the two ImageNets, a 40-epoch window for CIFAR-10/100 and TinyImageNet, frozen DINOv2 features for CUB-200), and m_0 is not comparable across instruments (§4.2). The tests are therefore read as coarse orderings, and decisive ordering contrasts are restricted to datasets within one instrument. The proxy-dynamics observations below are inconclusive about absolute boundary ordering.

1. **Total pixels per class** ($N \times \text{res}^2$). No direction fits. Within the 40-epoch-window group, CIFAR-100 has the fewest pixels per class and flips at ≈85 (grid bracket 50–100), TinyImageNet carries $4\times$ its pixels per class and stays on the LFrac side beyond 350, and CIFAR-10 carries $10\times$ yet flips at ≈ 222, below TinyImageNet's censoring bound: no monotone reading survives. The two ImageNets (10-epoch window) do not rescue one either, carrying over $100\times$ CIFAR-100's pixels per class yet flipping at $m_0 \approx 46$ and 66.
2. **DINOv2 per-class effective rank** (participation ratio of the per-class feature covariance; a uniform 25 samples per class because CUB-200 has about 30, so with centered features the estimator is capped at 24). Values: ImageNet-1k 13.4, ImageNet-100 13.6, CUB-200 13.7, CIFAR-100 14.7, TinyImageNet 15.1, CIFAR-10 16.0. Within the 40-epoch group the order is wrong: CIFAR-10 has the highest rank but a boundary of 222, below TinyImageNet's (> 350); the 10-epoch pair is ordered correctly, by 0.2. The absolute values are dominated by the sample count (at 50 and 100 samples per class, where the pool allows, they rise to 19.5–24.1 and 24.9–31.9), but the ordering of the five datasets with enough samples is identical at 25, 50, and 100, and so is the CIFAR-10/TinyImageNet reversal.
3. **Proxy training dynamics** (final accuracy, mean learnability score, easy-sample pool fraction). These statistics do not supply a validated mapping to absolute m_0. CUB-200 combines an easy-sample pool fraction of about 0.374 with final proxy training accuracy 0.895 and a high-score easy-sample fraction of only 0.026, illustrating their dependence on fitting dynamics. Its LFrac advantage extends across the measured fraction range, but the corresponding absolute constraint is only $m_0 \gtrsim 22$. This cannot order its boundary relative to the measured crossovers on other datasets, and is not used as a falsifying contrast. Further matched-protocol evaluation would be needed to assess these statistics as predictors.
4. **Proxy fitting difficulty** (training accuracy at epoch 10). Direction is opposite: ImageNet-100 is easier to fit (86% vs. ImageNet-1k's 48%), yet its m_0 is larger ($66 > 46$); CIFAR-10 and CIFAR-100 have nearly identical fitting difficulty (90% vs. 89%), yet their m_0 differ by $2.4\times$ (≈ 222 vs. $\approx$ 85–100).
5. **Image resolution**. It cannot distinguish ImageNet-1k (46) from ImageNet-100 (66), both at 224 pixels, nor CIFAR-10 (≈ 222) from CIFAR-100 ($\approx$ 85–100), both at 32 pixels.
6. **Inter-class confusion** (the perplexity of the validation-set confusion matrix after removing the diagonal and renormalizing, i.e. $\exp(\text{entropy})$; budget-aligned at 200 samples per class,

5-epoch probe). The reading depends on how the entropy is normalized for class count. Raw perplexity puts ImageNet-1k above ImageNet-100 (23.0 vs. 17.9) although its m_0 is smaller ($46 < 66$); entropy normalized by its maximum $\ln(C-1)$ reverses the order (0.45 vs. 0.63), because ImageNet-1k has ten times as many classes to be confused with. The descriptor gives no stable direction on the one contrast tested, and the test stopped there under the preregistered hard-stop rule.

F UNDERLYING NUMBERS FOR FIGURES 1–4

The reading rule for the crossover location is uniformly §3.4: scanning from the low-budget side, linearly interpolate between observed points and take the first zero crossing from positive to negative; the zero-crossing marks in Figure 1 follow the same rule (the ImageNet-1k mark 46 is the linear interpolation between the two points f=0.03–0.04, $K = 10$ arm).

F.1 Figure 1. No curve is a best-K envelope; each is one declared arm. The ImageNet-1k curve values are in the three-arm table of Appendix A (the curve uses the $K = 10$ arm); the ImageNet-100 curve is the full-pool row of F.2 (K=10, m-grid), whose visible m=90 back-crossing is counted toward transition-zone uncertainty under the §3.4 rule; the f=0.05–0.20 segment of the CIFAR-100 and TinyImageNet curves is the $K = 40$ arm of the §4.1 phenomenon group (Appendix A), and the f=0.50/0.70 segment is the wide-sweep native $K = 20$ arm (Appendix B). The CIFAR-10 curve is the $K = 40$ arm throughout: the low-budget extension of Appendix A at m=25–150 spliced with the phase-D grid at m=250–1000. The CUB-200 curve is the $K = 10$ arm against the self-run frozen-DINOv2 Herding (arm table in Appendix A). The CIFAR-100 dot is the first zero-crossing interpolation of its K=40 arm (≈ 85); the surrounding K=40 grid brackets the flip in 50–100.

F.2 Figures 2 and 3(a): ImageNet-100 pool shrinking, accuracy gap (LFrac − Herding). Numbers are below.

m	N=1267	N=600	N=300	N=150
25	+5.98	+5.84	+5.33	+4.73
50	+2.77	+2.07	+2.45	+1.43
63	+0.78	+0.60	+0.12	+0.16
75	-1.96	-0.21	-0.80	+0.21
90	+0.88	-2.47	-2.15	-0.37
130	-4.33	-3.11	-2.07	–
180	-6.29	-4.36	-2.24	–
m_0	66.4	71.9	64.6	80.4
$f_0 = m_0/N$	0.052	0.120	0.215	0.536

$N = 150$ only tests $m \leq 0.7N$; the removal of $m = 105$ is in Appendix C. Per-pool ceilings (full data, 3 seeds): 84.44 / 79.07 / 69.95 / 56.57.

Seed-level bootstrap ($B = 10{,}000$; each replicate redraws the 3 physical seeds with repetition, preserving the cross-budget correlation within a seed, and recomputes the interpolated crossover): 95% intervals are 66.4 [62.1, 69.4] / 71.9 [66.6, 77.6] / 64.6 [63.0, 64.9] / 80.4 [60.3, 88.1]. In 0 of 10,000 joint replicates did the four levels span a band wider than 30 samples per class, and no replicate was censored or degenerate. With 3 seeds the bootstrap distribution has exactly 10 support points per level, so these intervals are coarse; they bound seed noise, not a smooth continuum.

Sensitivity to the crossing rule. The §3.4 rule reads the first positive-to-negative crossing. A stricter diagnostic reads the last crossing after which no measured budget returns positive. The two rules differ only where a back-crossing exists: the full ImageNet-100 pool reads 96.8 instead of 66.4 (m=90 is the only positive point after the first flip), the three shrunken ImageNet-100 pools are unchanged, and the CIFAR-100 pools of F.4 read 57.2 / 61.5 / 56.7 / 47.6 instead of 43.2 / 39.5 / 42.6 / 36.3. The verdict does not depend on the rule. Fitting $\log m_0 = a + \gamma \log N$ across the four pools of a family, a fixed-fraction boundary predicts $\gamma = 1$ and a fixed per-class boundary $\gamma = 0$: the first-crossing readings give $\gamma = -0.07$ (ImageNet-100) and $+0.08$ (CIFAR-100), the

last-crossing readings $+0.10$ and $+0.10$. Under either rule the exponent is an order of magnitude below the fraction hypothesis; what the rule changes is the width of the band, 65–80 versus 65–97 on ImageNet-100.

F.2b Selection overlap under pool shrinking (convergence check). In a shrunken pool, any two class-balanced selectors choosing m of N per class overlap at the chance rate m/N in expectation even when they select independently (the combinatorial floor is $\max(0, 2m - N)/m$), and the chance rate reaches 0.6 at N=150, m=90. If this convergence of the candidate sets produced the sign flip, the flip would sit at a fixed overlap level, that is, at a fixed m/N. Measured same-seed overlap between the LFrac and Herding selections tracks the chance line with a small excess at every level and budget (full pool $+0.005$–0.028, N=600 $+0.007$–0.029, N=300 $+0.024$–0.036, N=150 $+0.024$–0.044; at N=150, m=75/90: observed 0.532/0.630 against chance 0.500/0.600). The two selectors keep disagreeing about the free $(1-m/N)$ remainder. At the crossover budgets themselves the overlap level spans 0.06 (full pool) to 0.46–0.53 (N=150) while m_0 stays in the 65–80 band: the data show a constant m, not the constant overlap a convergence artifact predicts. The CIFAR-100 family repeats the pattern (excess $+0.009$–0.036 across its four pools).

F.3 Figure 3(b): fixed-canvas bandwidth, accuracy gap (LFrac − Herding). Bold entries are the 3-seed bracket, the rest are single-seed exploratory points (Appendix B).

r	f0.01	f0.02	**f0.03**	**f0.05**	**f0.07**	**f0.10**	f0.20	f0.40	Ceiling
32	+4.66	+7.56	**+6.52**	**+2.40**	**-1.31**	**-3.16**	-4.34	-5.56	77.06±0.07
64	+4.34	+7.40	**+6.25**	**+1.57**	**-1.28**	**-4.88**	-6.42	-5.54	81.36±0.29
128	+5.92	+5.96	**+5.41**	**+1.17**	**-2.30**	**-4.03**	-6.58	-6.20	83.75±0.37
224	+3.54	+6.46	**+5.06**	**+0.27**	**-2.15**	**-4.43**	-6.78	-6.84	84.44±0.14

All four bandwidths flip between f=0.05 (positive) and f=0.07 (negative), roughly 63–89 samples per class.

Within that shared bracket the interpolated first-crossing readings do drift with resolution: $f_0 = 0.0522$ [0.0460, 0.0605] / 0.0568 [0.0499, 0.0665] / 0.0610 [0.0554, 0.0651] / 0.0629 [0.0610, 0.0641] for $r = 224/128/64/32$ (seed-level bootstrap as in F.2, 3-seed bracket only), or $m_0 \approx 66/72/77/80$ samples per class. The drift is systematic rather than seed noise: $f_0(r{=}32) > f_0(r{=}224)$ in every one of the 10,000 paired replicates, which with three seeds cover only ten distinct resample multisets per arm, so this is a coarse seed-level check rather than a fine significance level. Its direction is consistent with hypothesis (ii): blurrier images, later crossover. The $49\times$ pixel reduction moves the interpolated reading by about $1.2\times$. Since effective pixel count is not proportional to classification-relevant information, this comparison does not reject an information-based explanation.

F.4 §4.2 CIFAR-100 pool-shrinking replication, accuracy gap (LFrac − Herding, native R18 features). Numbers are below.

m	N=500	N=250	N=150	N=100
20	+4.17	+3.75	+4.45	+2.24
30	+3.73	+3.04	+2.69	+1.17
40	+2.14	-0.15	+0.69	-0.67
45	-1.17	-0.14	-0.63	+0.65
55	+1.58	-2.09	+0.62	-1.86
60	-2.02	+0.73	-1.21	-0.88
70	-1.10	-4.15	-0.15	-2.42
90	-2.77	-3.42	-2.71	–
120	-3.84	-4.58	–	–
m_0	43.2	39.5	42.6	36.3
$f_0 = m_0/N$	0.086	0.158	0.284	0.363

Within the transition neighborhood (about 40–60 per class) the sign is non-monotonic and Herding has large seed variance (e.g. ± 2.91 at $N = 500$, $m = 55$), so the main text reports a narrow band

and a transition neighborhood (§4.2). Seed-level bootstrap 95% intervals (as in F.2): 43.2 [41.4, 44.7] / 39.5 [37.9, 47.0] / 42.6 [36.7, 47.0] / 36.3 [28.3, 51.9]; the probability that the four levels span a band wider than 15 samples per class is 0.036, with no censored replicates. Fixed-m Uniform anti-artifact check: $m = 40$ is 34.4 / 34.3 / 33.5 / 33.7, $m = 55$ is 41.9 / 43.6 / 43.2 / 42.6, nearly constant across the four pools.

F.5 Figure 4. The values are the three columns $f{=}0.10/0.17/0.20$ of Construction A in Appendix D.